

# wKinMut-2: Identification and Interpretation of Pathogenic Variants in Human Protein Kinases

**Vazquez, Miguel; Pons, Tirso; Brunak, Søren; Valencia, Alfonso; Gonzalez-Izarzugaza, Jose Maria**

*Published in:*
Human Mutation

*Link to article, DOI:*
10.1002/humu.22914

*Publication date:*
2016

*Document Version*
Peer reviewed version

# wKinMut-2: Identification and interpretation of pathogenic variants in human protein kinases

Miguel Vazquez[1,§,*], Tirso Pons[1,§], Søren Brunak[2,3], Alfonso Valencia[1] and Jose M.G. Izarzugaza[3,*]

[1] Structural Biology and BioComputing Programme, Spanish National Cancer Research Centre (CNIO). Melchor Fernández Almagro, 3 28029 Madrid, Spain.

[2] Novo Nordisk Foundation Center for Protein Research, Faculty of Health Sciences, University of Copenhagen, Blegdamsvej 3A, 2200 Copenhagen, Denmark.

[3] Center for Biological Sequence Analysis (CBS), Systems Biology Department, Technical University of Denmark (DTU), Kemitorvet, building 208, 2800 Kgs. Lyngby, Denmark

**To whom correspondence should be addressed. (txema@cbs.dtu.dk)*

[§]*Contributed equally to this work.*

*Contact information*

Jose M. G. Izarzugaza: txema@cbs.dtu.dk

Miguel Vazquez: miguel.vazquez@cnio.es

*Grant sponsors*:

This work is supported by The Danish National Advanced Technology Foundation (The Genome Denmark platform, grant 019-2011-2) and by the EU FP7 project ASSET (grant agreement 259348).

## ABSTRACT

Most genomic alterations are tolerated while only a minor fraction disrupts molecular function sufficiently to drive disease. Protein kinases play a central biological function and the functional consequences of their variants are abundantly characterized. However, this heterogeneous information is often scattered across different sources, which makes the integrative analysis complex and laborious. wKinMut-2 constitutes a solution to facilitate the interpretation of the consequences of human protein kinase variation. Nine methods predict their pathogenicity, including a kinase-specific random forest approach. To understand the biological mechanisms causative of human diseases and cancer, information from pertinent reference knowledgebases and the literature is automatically mined, digested and homogenized. Variants are visualized in their structural contexts and residues affecting catalytic and drug-binding are identified. Known protein-protein interactions are reported. Altogether, this information is intended to assist the generation of new working hypothesis to be corroborated with ulterior experimental work. The wKinMut-2 system, along with a user manual and examples is freely accessible at http://kinmut2.bioinfo.cnio.es, the code for local installations at https://github.com/Rbbt-Workflows/KinMut2.

## KEYWORDS

Protein kinase, variants, pathogenicity prediction, variant annotation, functional impact

## INTRODUCTION

Only a minor fraction of the large number of variants discovered with current high-throughput next generation sequencing (NGS) methodologies (Sjöblom et al. 2006; Greenman et al. 2007; Wood et al. 2007) are causally implicated in disease. Discerning between disease-causing and neutral variants remains a challenge that requires computational methods to guide and prioritize the experiments (Baudot et al. 2009). The protein kinase superfamily plays a central role in the cell and its members have been studied in detail. Consequently, this superfamily constitutes a key example where abundant experimental evidence linking variants and disease, including some types of cancer, exists (Krallinger et al. 2009; Stratton et al. 2009). Albeit abundant respect to other protein families, this information is often heterogeneous, disperse and incomplete, which complicates the analysis of the consequences of variants.

wKinMut-2 represents an evolution on the original system that improves content and functionalities. Following our previous work (Izarzugaza et al. 2012; Izarzugaza et al. 2013), wKinMut-2 collects, integrates and digests relevant information extracted from diverse sources, including the one directly extracted from the literature. Residues are evaluated with respect to their catalytic role, including binding of known drugs broadly used in the clinic, and the involvement in interaction interfaces. Variants are characterized at the protein, domain and residue levels and represented within the context of 3D structures of the proteins. Furthermore, wKinMut-2 integrates information on predicted consequences of variants from 9 complementary methods, including a new random forest-based method trained with kinase specific information.

Previous attempts to the interpretation of variants affecting the human kinome exist (Supp. Table S1), although they often incorporate a limited number of information

sources and differ in rationale. For example, MoKCa (Richardson et al. 2009) constitutes a valuable resource that integrates information for known kinase variants. Annotations span different levels, consider the structural context of variants and include examples of expert curated assessment of the functional consequences. Unfortunately, annotations cannot be inferred on-the-fly for newly discovered genetic events; this static character hinders the applicability of MoKCa in a clinical setting. Examples of dynamic prediction servers also exist. Torkamani (Torkamani and Schork 2007) proposed a machine-learning methodology to predict the impact of kinase variants in cancer onset. Moreover, ProKinO (McSkimming et al. 2015) follows a new approach that results in the generation of testable hypotheses through an iterative process. This includes aggregate ontology querying and conceptualization of diverse forms of information (e.g., conserved sequence, structural motifs) in a machine-readable, human-understandable form. Both methodologies might identify causative variants but lack the contextualization of the findings with existing knowledge. Contrarily to these approaches, wKinMut-2 constitutes a one-stop shop to ease the interpretation of the consequences of variants in human protein kinases through the combination of imputed and validated information: Predictions of the functional consequences of kinase variants are integrated with information from knowledge bases, literature mining, and 3D analysis of protein structures.

## METHODS

### Web server framework

wKinMut-2 has been implemented primarily in Ruby as a workflow accessible through a REST interface. Consequently, results can be rendered in multiple formats including TSV (tab-separated values), JSON and HTML. Here we focus mainly on the latter, although fully programmatic access to the predictions is documented in the web site. External resources of information, such as gene descriptions or iHOP interactions, are queried remotely through the Internet on demand; subsequent accesses benefit from a cache system. Overall back-end performance is improved by a caching scheme that allows persisting job results and faster web interface visualization. wKinMut-2 is publicly available, including documentation and help pages, at http://kinmut2.bioinfo.cnio.es. The source code has been deposited in GitHub (https://github.com/Rbbt-Workflows/KinMut2) under a GPL version 3 licence.

### Submission of variant for analysis

Figure 1 displays a typical workflow in the analysis of protein kinase variation with wKinMut-2. Single point missense events affecting the human kinome are the input to the server. Variants are defined by a UniProt accession, a position in the protein and the wild-type and alternative amino acids. Consequently, a change from Valine to Glutamate in position 600 of the B-Raf proto-oncogene would be encoded as P15056 V600E (Supp. Fig. S1, panel a). Non-standard amino acids (B and Z) will be decomposed into separate instances of their standard counterparts (D, N and E, Q, respectively) whereas synonymous and truncating variants will be excluded from the analysis. Additionally, we exclude input instances where the introduced wild-type amino acid does not coincide with the expected equivalent position in the canonical

protein sequence. This identifies input errors and avoids incorrect annotations in the downstream analysis.

## Datasets and data pre-processing

In the present work we used publicly available datasets: (1) UniProt Variant Pages (Yip et al. 2008), (2) KinMutBase (Ortutay et al. 2005), (3) Kin-Driver (Simonetti et al. 2014), (4) COSMIC (Bamford et al. 2004), and (5) ClinVar (Landrum et al. 2014) . We decided to use all variants for which a classification as neutral/disease was available in UniProt Variant Pages. No filtering based on the disease that might have originated the classification was exerted. A total of 850 variants affecting 299 kinases were not used in the analysis because they were listed as unclassified in UniProt. After this pre-processing step, 1021 unique variants in 84 proteins kinases constitute the disease class whereas the neutral class consists of 2668 variants in 450 proteins. A link to the full list of selected variants is available in the website (see help pages). Other additional details and the distribution of variants respect to the classification features (i.e., membership to kinase groups, gene ontology terms, PFAM domains, amino acid and their biochemical properties, residue conservation, and functional annotations in UniProt, FireDB and Phospho.ELM) are displayed along with the description of each individual feature in Supp. Fig. S4 - S13.

## Data availability and reproducibility of results

The datasets utilised for training and evaluation of KinMutRF, included the data splits in the 10-fold cross-validation, are available through the help pages of our web server. The selected 3689 variants described in the previous section were randomly divided into 10 bins of similar sizes, constrained only by the rule that different

variants affecting the same protein should be forced into the same bin. We incorporated this rule to avoid overestimating the performance of the classification due to similarities between elements of the training and evaluation sets.

For each dataset, we compiled tab-separated files containing the classification features. To assess the contribution of individual features we calculated their information gain with respect to the classification classes using the InfoGainAttributeEval function in Weka. More details about identification of the most relevant features for prediction (see Supp. Table S2), and evaluation of KinMutRF's performance according to external datasets are provided in Supplementary Materials.

## RESULTS

### Interpretation of the consequences of variants

By default, wKinMut-2 presents a 'Summary table' that provides a quick overview of the results and helps to prioritize variants of interest for detailed analysis. The summary highlights the trait/disease annotations, and molecular details about the investigated variants. For example: i) relationships among human variations and phenotypes with supporting evidences annotated in ClinVar (Landrum et al. 2014); ii) number of COSMIC samples with variants overlapping that same residue; iii) variants overlapping a post-translational modified residue, a ligand-binding or catalytic residue, and residues experimentally altered by mutagenesis; iv) number of other damage predictors that predict the variant as damaging; and v) the groups to which the kinases belong according to the classification in KinBase (Manning et al. 2002a; Manning et al. 2002b; Miranda-Saavedra and Barton 2007), and the color coded pathogenicity scores according to our kinase-specific random forest-based predictor. Further details will be provided in the corresponding section. Note that a summary

report of the prediction and all relevant files for the prediction are available for direct download from this page.

The primary goal of wKinMut-2 is to aid computational biologists and clinicians to understand and to interpret the consequences of disease-causing variants acquired by human protein kinases. Expanded information is gathered and provided when the user clicks on the 'View details' button (Supp. Fig. S1, panel b). This spawns a multi-tabular page including the characterization of the variants at the protein, domain and residue level; the representation of the variants onto known protein structures and models including the analysis of the disruption of the protein-protein interaction interfaces; the prediction of pathogenicity using different methodologies; the collection of mentions of the variants in dedicated knowledgebases, complemented with information mined from the literature and the study of the known and predicted interaction partners. Each of these tabs will be detailed in the following sections and their main characteristics explained.

**General description of the mutated kinases**

The first view presents a 'General' tab (Supp. Fig. S1, panel c). This section compiles diverse information about the kinase that harbors a variant of interest. This includes the gene name and the description from UniProt, the protein identifier in Ensembl and the classification in kinase groups as defined by KinBase (Manning et al. 2002a; Manning et al. 2002b; Miranda-Saavedra and Barton 2007). In addition, as a proxy to understand the cellular role of the protein, we list GeneOntology (Ashburner et al. 2000) annotations grouped by subontology (i.e., Molecular Function, Cellular Compartment and Biological Process). Besides we include three main types of information about: i) essential or non-essential phenotype-changing of the homologous gene in mouse based on the information collected by dbNSFP (Liu et

al. 2013) from the Mouse Genome Informatics database (Georgi et al. 2013); ii) US FDA (Jänne et al. 2009) approved protein kinase inhibitors (http://www.brimr.org/PKI/PKIs.htm) and iii) bioactive compounds and screening data for kinases, extracted from Kinase SARfari an integrated chemogenomics workbench available at EMBL-EBI (https://www.ebi.ac.uk/chembl/sarfari/kinasesarfari).

**Mapping variants onto protein structures**

The effect of variants is better understood in their own structural contexts under the assumption that pathogenic variants distort protein function and structural stability (Izarzugaza et al. 2009b; Izarzugaza et al. 2011). Mapping variants from sequences to 3D structures is not trivial (Izarzugaza et al. 2009a). To bridge this gap, wKinMut-2 enables the visualization of variants onto the available structures and models (Supp. Fig. S1, panel d) in PDB and Interactome3D. We chose Jmol provided the versatility of the built-in console, which permits customized representations. In addition, information contextualizing the variant regarding to the affected protein domains is retrieved from Ensembl.

**Predicting the pathogenic potential of variants**

wKinMut-2 provides the likelihood of a given variant to be causally implicated in disease in the 'Pathogenicity' section (Supp. Fig. S1, panel f). We provide predictions from nine different methodologies ranging from the most classical evolutionary perspective provided by SIFT, to a new random forest machine learning classifier adapted for the prediction of pathogenesis within the human kinome. Predictions are directly extracted from dbNSFP v2.8 (https://sites.google.com/site/jpopgen/dbNSFP) (Liu et al. 2013) and include likelihood scores from SIFT release August 2011 (http://sift.jcvi.org/) (Ng and Henikoff 2003), Polyphen2 v2.2.2

(http://genetics.bwh.harvard.edu/pph2/) (Adzhubei et al. 2010), LRT release November 2009 (http://www.genetics.wustl.edu/jflab/lrt_query.html) (Chun and Fay 2009), MutationTaster release 2013 (http://www.mutationtaster.org/) (Schwarz et al. 2010) , MutationAssessor release 2 (http://mutationassessor.org/) (Reva et al. 2011), FATHMM v2.3 (http://fathmm.biocompute.org.uk) (Shihab et al. 2013), VEST3 v3.0 (http://karchinlab.org/apps/appVest.html) (Carter et al. 2009) and CADD v1.0 (http://cadd.gs.washington.edu/) (Grimm et al. 2015). In addition to these, wKinMut-2 implements a new method specific to the protein kinase superfamily named KinMutRF. This method relies on a random forest classifier consisting of 26 decision trees that evaluate a number of sequence-derived features that characterize variants affecting human protein kinases at different levels: a) at the gene level, including membership to a Kinbase group and Gene Ontology categories; b) at the domain level, using PFAM domains; and c) at the residue level, involved amino acids types, changes in biochemical properties, functional annotations from UniProt (Yip et al. 2008), Phospho.ELM (Dinkel et al. 2011) and FireDB (Lopez et al. 2007). These are described in detail in the corresponding section of the Supplementary Materials. The focus on the protein kinase superfamily enabled the choice of features unique to this superfamily. These kinase-specific annotations greatly increase the accuracy of the classification of variants. Each prediction is accompanied with a reliability score and the annotations giving rise to the predictions to facilitate the biological interpretation of the results. To provide an example of the functionality of wKinMut-2, we applied the method to a reduced dataset of BTK variants. Table 1 shows an output of the wKinMut-2 annotations in comparison with the PON-BTK method. The evaluation of the prediction performance of the 9 methods in different kinase datasets following

243 the best practices in the field (Vihinen 2012; Grimm et al. 2015) is given in Supp.
244 Table S3.

245 In the following sections we will discuss how the combination of prediction of
246 pathogenicity and integration of information can enhance our understanding of the
247 consequences of variants and facilitate the generation of working hypotheses.

248

249 **Variation in protein interaction interfaces and in the vicinity of the known**
250 **functional sites**

251 Structure-PPi (http://structureppi.bioinfo.cnio.es/Structure) (Vazquez et al. 2015) is a
252 system to facilitate the analysis of variants in the context of protein complexes. The
253 system combines information from protein structures with functional annotations
254 from a number of relevant databases and reports protein features (e.g., functional
255 domains, known somatic variation in different types of cancer, UniProt annotations
256 from missense variants, ligand binding residues, catalytic sites) that overlap the
257 variant's 'direct matches' or their 'neighbors' in close physical proximity (Supp. Fig.
258 S1, panel e) defined as 5 angstroms spatial distance or adjacent in the sequence if no
259 PDB covers that area. When variants affect the interfaces of protein complexes (when
260 the variant is at a distance of less than 8 angstroms from a residue in the partner
261 protein), Structure-PPi also reports the partner proteins, and the residues in those
262 proteins that are in close proximity to the variant.

263 In order to illustrate the power of the integration of information in the analysis of the
264 consequences of variation, we used Structure-PPi to analyze the variants that
265 KinMutRF failed to predict as pathogenic in the Kin-Driver (Simonetti et al. 2014)
266 collections, as described in the Supplementary Material. Structure-PPi identified
267 (features score>3) 64.89% (61/94) and 61.54% (8/13), respectively, of the activating

and deactivating variants erroneously predicted by the pathogenicity classifier. Complementary annotations from Structure-PPi provide information about the functional role of the affected residues. For example, considering a deactivating variant in the STK11 serine/threonine-protein kinase p.D176N, Structure-PPi reveals that Asp-176 is a catalytic residue (Firestar and UniProt features) that accommodates several variants in lung cancer. This is also true for his structural neighbors, which present variants in different cancer types including lung, colon, cervix, kidney and leukemia. Two of these structural neighbors Lys-178 and Asn-181 are located in the ATP binding pocket; being in close structural proximity, Asp-176 might be relevant for the correct positioning of them.

**Variation in scientific knowledgebases**

A plethora of scientific databases approaches the consequences of variants from a variety of complementary perspectives. wKinMut-2 (Supp. Fig. S1, panel g) aggregates information from the UniProt Variant Pages (Yip et al. 2008), KinMutBase (Ortutay et al. 2005), Kin-Driver (Simonetti et al. 2014), COSMIC (Bamford et al. 2004) and ClinVar (Landrum et al. 2014) and provides digested contextual information about the variants. Of particular interest, experimental evidence relating variants and disease, being cancer a recurrent example. Users of previous versions of the tool would notice that SAAPdb is not included in the current implementation, as its authors have discontinued it.

**Mining of variants from the literature**

Mining the literature complements the information sourced from the databases. A substantial number of references describe the experimental conditions or the specific background of the patients. Furthermore, in some cases the mined articles may describe the biochemical mechanisms by which variants might lead to the

corresponding diseases. As these aspects are essential for the correct interpretation of the consequences of variation, wKinMut-2 collects mentions to kinase variants from the literature using SNP2L (Krallinger et al. 2009). Our literature-mining pipeline extracts and disambiguates automatically references to variants in both PubMed abstracts and full text articles. Strict internal filters increase the reliability of SNP2L; after variation mentions and kinases have been identified in a text, SNP2L checks the coherence of the mined amino acids and the expected sequences to reduce the number of spurious hits. SNP2L results are displayed under the 'Literature' tab of wKinMut-2 (Supp. Fig. S1, panel h).

**Identification of interaction partners**

Two methods are used for the exploration of protein-protein interactions. First, the 'iHop' tab (Supp. Fig. S1, panel i) elucidates known and inferred protein-protein interaction partners. iHop (Hoffmann and Valencia 2005) has proven a useful text mining system to automatically mine interaction partners in PubMed abstracts. In order to provide the interactions in their context, results displayed include specific sentences from the literature as well as pointers to the original articles. Secondly, the 'String' tab (Supp. Fig. S1, panel j) graphically shows the interactions described for the human kinome in the homonymous database (Szklarczyk et al. 2015). Evidence for interactions include gene neighborhood and gene fusion events, co-occurrence of the proteins across species, co-expression of genes in the same or other related species, experiments stored in databases as well as text-mining of PubMed abstracts.

**USE CASES**

To illustrate the utility of wKinMut-2, we describe next two examples where the system is used to draw hypotheses about the role of kinase variants in the onset of human diseases.

**The p.V600E variant in the B-Raf proto-oncogene.**

The B-Raf proto-oncogene is involved in the transduction of the mitogenic signals from the cell to the membrane of the nucleus. As described in the 'General' tab of wKinMut, this kinase regulates the MAPK/ERKs signal transduction pathway by phosphorylation of MAP2K1. The variant in position 600 from Valine to Glutamate used to illustrate the example in Supp. Fig. S1 is recurrently found in the literature associated to certain types of cancer including colorectal cancer, sarcoma, melanoma, thyroid carcinoma and ovarian serous carcinoma. These observations are coherent with the information obtained from the 'Literature' and 'Databases' tabs of wKinMut-2. Furthermore, the 'Pathogenicity' tab shows that there is a consensus among the predictors to classify this variant as pathogenic. According to Structure-PPi annotations, p.V600E could affect the ATP-binding because due to the proximity to Phe-468, localized in the ATP-binding region (PDB ID: 3tv4, chain B). Phe-468 has been reported as substituted (p.F468S, VAR_035097) in Cardiofaciocutaneous syndrome 1 as well as in colorectal cancer. In addition, a characterization of the functional and biochemical properties is provided. Although p.V600E is harbored in the protein kinase domain, the affected residue does not seem to play a direct functional role. By contrast, the p.V600E variant may confer the activated phenotype by destabilization of the structure of the protein as it induces substantial changes in biochemical properties like hydrophobicity, formal charge, volume and C-beta branching. In addition, several records in the literature from the 'Literature' and

‘iHop’ tabs suggest that p.V600E might be associated with genomic instability leading to secondary genetic events that might account for its aggressive phenotype.

**The p.P250R variant in the interface between the human fibroblast growth factor receptor 3 and its interaction partner FGF1.**

This example illustrates a case where the predictors do not provide a consensus classification. The p.P250R variant is considered pathogenic by KinMutRF, SIFT, Polyphen2 and FATHMM, whereas MutationTaster, MutationAssessor, VEST3 and CADD consider it neutral. These predictions are inconclusive. Interestingly, the variant appears in the ‘Databases’ tab associated with Muenke’s syndrome and the ‘Literature’ tab reports plenty of pointers to publications discussing the role of this variant in some aberrant phenotypes including craniosynostosis, epidermal hyperplasia and Apert and Crouzon syndromes. Moreover, Proline-250 is adjacent to an activating variant, p.S249C, previously described in benign epidermal tumors (Logié et al. 2005).

The correct assembly of two of more proteins in a complex is highly dependent on the physic-chemical properties of the interaction interfaces. For this reason, the amino acids conforming those functional surfaces are tightly preserved through evolution and modification of such residues commonly leads to aberrant phenotypes. According to the Structure-PPi, both p.P250R and p.S249C target the interaction interface between FGFR3 and FGF1 (Supp. Fig. S2). Following the diverse literature pointers provided, we hypothesize that the stimulation of downstream developmental pathways might cause the aberrant phenotype and we propose the disruption of the protein-protein interfaces as a plausible mechanism to be confirmed by further experimental work.

## DISCUSSION

Protein kinases constitute a broad superfamily involved in relevant physiological functions. Although most variation affecting protein kinases is functionally neutral and tolerated by the cells, many examples have been published associating human protein kinase variants with disease, and particularly with cancer. With the advent of NGS technologies, the identification of variants is routine in most laboratories; however, the exploration of the consequences on the phenotypes of each individual event remains a considerable challenge.

Here we describe a unified system, named wKinMut-2, that facilitates the identification and the interpretation of pathogenic variants in the human kinome.

The system summarizes information about pathogenicity of variants annotated in the ClinVar database, number of tumor samples with variants over that same residue or close to that residue, number of sites that have been experimentally altered by mutagenesis, sites of post-translational modifications, ligand-binding sites, etc.

wKinMut-2 offers direct prediction of the potential pathogenicity of non-synonymous single amino acid variants with eight independent methods plus a new kinase-specific predictor. As most pathogenic variants disrupt the function and structural stability of the protein, a random forest-based approach evaluates changes in the properties of the affected amino acids. The system was trained with features that include the physicochemical properties of the amino acids, the sequence conservation according to SIFT and functional annotations from Phospho.ELM, FireDB and UniProt. These features apply generally to any protein. In addition, we consider kinase-specific properties such as membership to a specific group of kinases, prevalence of certain GO terms in disease associated kinases and the relevance of PFAM domains. One could argue the need for a new family-specific prediction

388 method, in the wealth of methodologies available. We performed a cross-validation
389 experiment that analyzed the prediction of 8 systems on the 3689 kinase variants in
390 UniProt for which an annotation of pathogenicity is available. The benchmark (Supp.
391 Table S3) shows that KinMutRF outperforms most classifiers, with the exception of
392 VEST3 where closely similar performance is achieved. It is important to highlight at
393 this point that the performance of predictors competing with KinMutRF might be
394 optimistically interpreted due to the fact that a fraction of variants in the evaluation
395 set used in this benchmark might had been presented to the classifier during its own
396 training phase. Some authors refer to this relevant effect as type I circularity (Grimm
397 et al. 2015). This effect has been considered in the development and evaluation of our
398 methodology. In addition to attaining full comprehension of the datasets, a
399 concomitant added value of an in-house system is the control over the frequency of
400 re-training. Particularly, in the event of new kinase variants being discovered and
401 classified. Moreover, VEST3 is intended for the identification of variation in exomes
402 rather than for the investigation of individual events. Consequently, it is designed to
403 partake in bioinformatics pipelines, often in memory-demanding computational
404 scenarios and no web interface is provided. This might not be optimal in a clinical
405 setting. Finally, our experience in the context of clinical cancer genomics suggests
406 that the combination of different prediction methods takes advantage of the different
407 prediction features and facilitates interpretation of the results in combination with
408 other available sources, as in the *FGF1* example presented above.
409 A common approach to exploit our integrative approach presented here would be to
410 start by using one or more of the scores from the pathogenicity predictors to prioritise
411 the variants of interests, and then proceed to an in-depth analysis using the other
412 integrated tools to devise a functional role and or a biochemical mechanism. Our

analyses demonstrated with an independent manually curated dataset that current methods to predict the pathogenicity of variants are not sensitive enough to characterize variants alone and that activating variation constitutes a particularly challenging scenario. This highlights the need of integrative tools that combine prediction and information compilation. wKinMut-2 integrates digested information from a number of selected resources, which helps to interpret the consequences of variants and can be used to draw hypothesis regarding the mechanisms linking variants to the alterations of the structure and function of the kinases, and indirectly point to the possible relation with observed phenotypes or diseases. The information provided, might constitute a basis for the design of new experimental work. Particularly relevant for the interpretation of the variants is the visualization of the information in the context of the corresponding protein structures, the mentions of the involvement of the variants in previous disease studies in publications or large-scale cancer genome projects. wKinMut-2 also includes information automatically extracted from the literature with our in-house tools; possible interaction partners are mined from PubMed abstracts with iHop and variant mentions are gathered from the literature using SNP2L, which provides a substantial addition to the information provided by public databases and repositories. Both tools display the particular sentences in the literature that support the association and supply external references to the original publications.

We have described hereby a tool to study the links between kinase variation and disease. The examples provided highlight how raw predictions of pathogenicity are complemented by the integration and digestion of information from multiple sources and how this information can be used to propose further experimental work. Nevertheless, it is still necessary that resources of the like develop further, aiming to

the automatic integration and processing of the wealth of information. The ultimate goal should be the reliable generation of hypotheses to explain the underlying biological mechanisms relating causative variants and diseases as a necessary step towards routine personalized treatment in the clinic.

## ACKNOWLEDGEMENTS

The authors are extremely grateful to the anonymous referees of this manuscript for their very pertinent comments and suggestions.

*Funding*: This work is supported by The Danish National Advanced Technology Foundation (The Genome Denmark platform, grant 019-2011-2) and by the EU FP7 project ASSET (grant agreement 259348).

*Author contributions:* All authors contributed in various degrees to the conceptual definition of the experiment. JMGI trained the classifier. MV and JMGI designed and implemented the web server. MV implemented the web services. All authors wrote, read and approved the final manuscript.

*Conflict of interest:* None declared.

## REFERENCES

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. 2010. A method and server for predicting damaging missense mutations. Nat. Methods 7: 248–249.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, et al. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat. Genet. 25: 25–29.

Bamford S, Dawson E, Forbes S, Clements J, Pettett R, Dogan A, Flanagan A, Teague J, Futreal PA, Stratton MR, Wooster R. 2004. The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website. Br. J. Cancer 91: 355–358.

Baudot A, Real FX, Izarzugaza JMG, Valencia A. 2009. From cancer genomes to

cancer models: bridging the gaps. EMBO Rep. 10: 359–366.

Carter H, Chen S, Isik L, Tyekucheva S, Velculescu VE, Kinzler KW, Vogelstein B, Karchin R. 2009. Cancer-specific high-throughput annotation of somatic mutations: computational prediction of driver missense mutations. Cancer Res. 69: 6660–6667.

Chun S, Fay JC. 2009. Identification of deleterious mutations within three human genomes. Genome Res. 19: 1553–1561.

Dinkel H, Chica C, Via A, Gould CM, Jensen LJ, Gibson TJ, Diella F. 2011. Phospho.ELM: a database of phosphorylation sites--update 2011. Nucleic Acids Res. 39: D261–7.

Georgi B, Voight BF, Bućan M. 2013. From mouse to human: evolutionary genomics analysis of human orthologs of essential genes. PLoS Genet. 9: e1003484.

Greenman C, Stephens P, Smith R, Dalgliesh GL, Hunter C, Bignell G, Davies H, Teague J, Butler A, Stevens C, Edkins S, O'Meara S, et al. 2007. Patterns of somatic mutation in human cancer genomes. Nature 446: 153–158.

Grimm DG, Azencott C-A, Aicheler F, Gieraths U, MacArthur DG, Samocha KE, Cooper DN, Stenson PD, Daly MJ, Smoller JW, Duncan LE, Borgwardt KM. 2015. The evaluation of tools used to predict the impact of missense variants is hindered by two types of circularity. Hum. Mutat. 36: 513–523.

Hoffmann R, Valencia A. 2005. Implementing the iHOP concept for navigation of biomedical literature. Bioinformatics 21 Suppl 2: ii252–8.

Izarzugaza JMG, Baresic A, McMillan LEM, Yeats C, Clegg AB, Orengo CA, Martin ACR, Valencia A. 2009a. An integrated approach to the interpretation of single amino acid polymorphisms within the framework of CATH and Gene3D. BMC Bioinformatics 10 Suppl 8: S5.

Izarzugaza JMG, del Pozo A, Vazquez M, Valencia A. 2012. Prioritization of pathogenic mutations in the protein kinase superfamily. BMC Genomics 13 Suppl 4: S3.

Izarzugaza JMG, Hopcroft LEM, Baresic A, Orengo CA, Martin ACR, Valencia A. 2011. Characterization of pathogenic germline mutations in human protein kinases. BMC Bioinformatics 12 Suppl 4: S1.

Izarzugaza JMG, Redfern OC, Orengo CA, Valencia A. 2009b. Cancer-associated mutations are preferentially distributed in protein kinase functional sites. Proteins 77: 892–903.

Izarzugaza JMG, Vazquez M, del Pozo A, Valencia A. 2013. wKinMut: an integrated tool for the analysis and interpretation of mutations in human protein kinases. BMC Bioinformatics 14: 345.

Jänne PA, Gray N, Settleman J. 2009. Factors underlying sensitivity of cancers to small-molecule kinase inhibitors. Nat Rev Drug Discov 8: 709–723.

503 Krallinger M, Izarzugaza JMG, Rodriguez-Penagos C, Valencia A. 2009. Extraction
504 of human kinase mutations from literature, databases and genotyping studies. BMC
505 Bioinformatics 10 Suppl 8: S1.

506 Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, Maglott DR.
507 2014. ClinVar: public archive of relationships among sequence variation and human
508 phenotype. Nucleic Acids Res. 42: D980–5.

509 Liu X, Jian X, Boerwinkle E. 2013. dbNSFP v2.0: a database of human non-
510 synonymous SNVs and their functional predictions and annotations. Hum. Mutat. 34:
511 E2393–402.

512 Logié A, Dunois-Lardé C, Rosty C, Levrel O, Blanche M, Ribeiro A, Gasc J-M,
513 Jorcano J, Werner S, Sastre-Garau X, Thiery JP, Radvanyi F. 2005. Activating
514 mutations of the tyrosine kinase receptor FGFR3 are associated with benign skin
515 tumors in mice and humans. Hum. Mol. Genet. 14: 1153–1160.

516 Lopez G, Valencia A, Tress M. 2007. FireDB--a database of functionally important
517 residues from proteins of known structure. Nucleic Acids Res. 35: D219–23.

518 Manning G, Plowman GD, Hunter T, Sudarsanam S. 2002a. Evolution of protein
519 kinase signaling from yeast to man. Trends Biochem. Sci. 27: 514–520.

520 Manning G, Whyte DB, Martinez R, Hunter T, Sudarsanam S. 2002b. The protein
521 kinase complement of the human genome. Science 298: 1912–1934.

522 McSkimming DI, Dastgheib S, Talevich E, Narayanan A, Katiyar S, Taylor SS,
523 Kochut K, Kannan N. 2015. ProKinO: a unified resource for mining the cancer
524 kinome. Hum. Mutat. 36: 175–186.

525 Miranda-Saavedra D, Barton GJ. 2007. Classification and functional annotation of
526 eukaryotic protein kinases. Proteins 68: 893–914.

527 Ng PC, Henikoff S. 2003. SIFT: Predicting amino acid changes that affect protein
528 function. Nucleic Acids Res. 31: 3812–3814.

529 Ortutay C, Väliaho J, Stenberg K, Vihinen M. 2005. KinMutBase: a registry of
530 disease-causing mutations in protein kinase domains. Hum. Mutat. 25: 435–442.

531 Reva B, Antipin Y, Sander C. 2011. Predicting the functional impact of protein
532 mutations: application to cancer genomics. Nucleic Acids Res. 39: e118.

533 Richardson CJ, Gao Q, Mitsopoulous C, Zvelebil M, Pearl LH, Pearl FMG. 2009.
534 MoKCa database--mutations of kinases in cancer. Nucleic Acids Res. 37: D824–31.

535 Schwarz JM, Rödelsperger C, Schuelke M, Seelow D. 2010. MutationTaster evaluates
536 disease-causing potential of sequence alterations. Nat. Methods 7: 575–576.

537 Shihab HA, Gough J, Cooper DN, Stenson PD, Barker GLA, Edwards KJ, Day INM,
538 Gaunt TR. 2013. Predicting the functional, molecular, and phenotypic consequences
539 of amino acid substitutions using hidden Markov models. Hum. Mutat. 34: 57–65.

Simonetti FL, Tornador C, Nabau-Moretó N, Molina-Vila MA, Marino-Buslje C. 2014. Kin-Driver: a database of driver mutations in protein kinases. Database (Oxford) 2014: bau104.

Sjöblom T, Jones S, Wood LD, Parsons DW, Lin J, Barber TD, Mandelker D, Leary RJ, Ptak J, Silliman N, Szabo S, Buckhaults P, et al. 2006. The consensus coding sequences of human breast and colorectal cancers. Science 314: 268–274.

Stratton MR, Campbell PJ, Futreal PA. 2009. The cancer genome. Nature 458: 719–724.

Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, Simonovic M, Roth A, Santos A, Tsafou KP, Kuhn M, Bork P, et al. 2015. STRING v10: protein-protein interaction networks, integrated over the tree of life. Nucleic Acids Res. 43: D447–52.

Torkamani A, Schork NJ. 2007. Accurate prediction of deleterious protein kinase polymorphisms. Bioinformatics 23: 2918–2925.

Vazquez M, Valencia A, Pons T. 2015. Structure-PPi: a module for the annotation of cancer-related single-nucleotide variants at protein-protein interfaces. Bioinformatics 31: 2397–2399.

Väliaho J, Faisal I, Ortutay C, Smith CIE, Vihinen M. 2015. Characterization of all possible single-nucleotide change caused amino acid substitutions in the kinase domain of Bruton tyrosine kinase. Hum. Mutat. 36: 638–647.

Vihinen M. 2012. How to evaluate performance of prediction methods? Measures and their interpretation in variation effect analysis. BMC Genomics 13 Suppl 4: S2.

Wood LD, Parsons DW, Jones S, Lin J, Sjöblom T, Leary RJ, Shen D, Boca SM, Barber T, Ptak J, Silliman N, Szabo S, et al. 2007. The genomic landscapes of human breast and colorectal cancers. Science 318: 1108–1113.

Yip YL, Famiglietti M, Gos A, Duek PD, David FPA, Gateau A, Bairoch A. 2008. Annotating single amino acid polymorphisms in the UniProt/Swiss-Prot knowledgebase. Hum. Mutat. 29: 361–366.

## FIGURE LEGENDS

Fig. 1: The different components of wKinMut-2 help to identify and interpret pathogenic variants in human protein kinases.

## TABLES

**Table 1**: Annotation of BTK variants by wKinMut-2 and PON-BTK methods.

| Variants | wKinMut-2 | PON-P | UniProt annotations[a] |
|---|---|---|---|
| p.E41K | disease | n.a. | No effect on phosphorylation of GTF2I |
| p.P189A | disease | n.a. | No effect on phosphorylation of GTF2I |
| p.Y223F | neutral | n.a. | Loss of phosphorylation of GTF2I |
| p.W251L | disease | n.a. | No effect on phosphorylation of GTF2I |
| p.R307G | disease | n.a. | Variant in XLA (VAR_006231); loss of activity |
| p.R307K | disease | n.a. | Loss of phosphorylation of GTF2I |
| p.K430E | disease | Path. | Variant in XLA (VAR_006242); loss of phosphorylation of GTF2I |
| p.R520Q | disease | Path. | Variant in XLA (VAR_006251); severe; prevents activation due to absence of contact between the catalytic loop and the regulatory phosphorylated residue |
| p.R525Q | disease | Path. | Variant in XLA (VAR_006255); severe; disturbs ATP-binding |
| p.Y551F | disease | Path. | Loss of phosphorylation of GTF2I |
| p.R562P | disease | Path. | Variant in XLA (VAR_006259). Corresponds to variant rs28935176:R562P |
| p.E567K | disease | Path. | Variant in XLA (VAR_006261); severe |
| p.E589G | disease | Path. | Variant in XLA (VAR_006265); moderate; interferes with substrate binding |
| p.G594E | disease | Path. | Variant in XLA (VAR_006268); interferes with substrate binding |
| p.G613D | disease | Path. | Variant in XLA (VAR_006272); interferes with substrate binding and/or domain interactions |
| p.Y617E | disease | n.a. | Defective in mediating calcium response |

[a]Identification codes in parenthesis corresponds to entries in the file humsavar.txt (release 2015_06 of 27-May-2015) from the Swiss-Prot Variant Pages available at http://www.uniprot.org/docs/humsavar. XLA: X-linked agammaglobulinemia. Annotations according to wKinMut-2. PON-P: prediction of pathogenicity by the PON-BTK method (Izarzugaza et al. 2009b; Izarzugaza et al. 2011; Väliaho et al. 2015). Predictions of pathogenic (Path.) or non-pathogenic (Non-Path) variants are available at http://structure.bmc.lu.se/PON-BTK/. n.a.: no annotation.